# Descriptive analytics

1. what is the distribution of number of cancelled trains across all train stations?

Justification

Answering this question is important as it helps to identify the frequency and extent of train cancellations across different stations. Understanding the distribution of cancelled trains is crucial for the SNCF to optimize their resources, reduce losses and improve customer satisfaction. This analysis can help identify the stations with the highest number of cancellations and the reasons for the cancellations, such as external causes or internal factors such as operational or technical difficulties.

1. What is the average trip times for trains departing from each train station?

Justification

This question is important as it helps to understand the average travel time for trains departing from each station. It can help identify the stations with longer travel times and the factors that contribute to the longer trip times, such as congestion or infrastructure issues. By identifying these factors, the SNCF can take appropriate measures to improve the service quality by reducing travel times and enhancing customer satisfaction.

1. How does the number of train trips departing from each station vary by month?

Justification

Answering this question is important as it helps to understand the seasonality of train trips departing from each station. By identifying the months with higher or lower demand, the SNCF can optimize their resources, such as trains and staff, to meet the demand. This analysis can also help the SNCF to identify the stations with the highest or lowest demand, and the factors contributing to the demand, such as holidays or events. By understanding the seasonality of train trips, the SNCF can take appropriate measures to improve their service quality

# Diagnostic

1. Are there any train stations that have a significantly higher rate of cancellations than others? If so what factors might contribute to this.

Justification

Answering this question is crucial for the SNCF as it would help them identify which stations need more attention and resources to improve their service quality and reduce cancellations. Additionally, identifying the factors contributing to the higher cancellation rate can help the SNCF take preventive measures to reduce cancellations in the future. This could potentially improve customer satisfaction and loyalty while also minimizing financial losses for the SNCF.

1. Do certain train routes tend to have a higher rate of delays than others across all train stations?

Justification

Answering this question would provide insights into the specific routes that are more prone to delays, allowing the SNCF to take proactive measures to reduce delays and improve overall service quality. This could potentially enhance customer satisfaction and loyalty, which would have a positive impact on the SNCF's reputation and financial performance. Additionally, identifying the root causes of delays could lead to the implementation of targeted interventions to address the issue and ultimately improve the reliability of the train service.

# Predictive

1. Can we predict the likelihood of train cancellations at the various stations based on historical data and external factors?

Justification

To minimize the impact of train cancellations, the SNCF can leverage historical data and external factors to develop a predictive model. By analyzing data on past cancellations, the SNCF can identify patterns and trends that may indicate the likelihood of future cancellations. For example, they can identify stations that have a higher rate of cancellations and investigate the root causes of these cancellations, such as weather-related disruptions or infrastructure issues. This analysis can help the SNCF make informed decisions about rescheduling trains, adding additional resources to stations, or communicating with passengers to minimize disruptions.

1. Can we predict which train routes are likely to experience delays based on historical data?

Justification:

Predicting which train routes or schedules are likely to experience delays can help guide targeted efforts to reduce delays and improve the passenger experience.

1. Can we forecast the average delay time at each station for the next month based on historical data?

Justification

Forecasting average delay times can help prepare passengers for potential delays and guide efforts to reduce delays and improve the overall performance of the railway system.

# PART 2

|  |  |  |  |
| --- | --- | --- | --- |
| **Questions** | **Variables** | **Statistical techniques** | **Why the statistical technique used** |
| 1.What is the distribution of number of cancelled trains | num\_of\_cancelled\_trains  Departure\_station | Groupby count () function along with n() | The groupby() function takes the num\_of\_cancelled\_trains and  Departure\_station as argument summarise() uses n() function to find count of items sold. |
| 2.What is the average trip times for trains departing from each train station? | Departure\_station  journey\_time\_avg | Groupby() function along with mean() | The groupby() function takes the departure station and journey time average as argument summarise() uses mean() function to find mean trip time of each station |
| 3. How does the number of train trips departing from each station vary by month? | Departure\_station  Month  total\_num\_trips | Groupby () function along with max()  min() | The groupby () function takes Departure\_station  Month  total\_num\_trips and summarises using sum() function to find stations with max number and min number of trips |
| 4.Are there any train stations that have a significantly higher rate of cancellations than others? If so what factors might contribute to this. | departure\_station  num\_cancelled\_trips | Anova test | Anova statistical technique would be a hypothesis test for comparing the means of multiple groups, such as one-way ANOVA (Analysis of Variance). This test allows us to assess whether there are significant differences in the mean number of canceled trains between different train stations. |
| 5.Do certain train routes tend to have a higher rate of delays than others across all train stations? | depature\_station  arrival\_station  av\_delay\_all\_departing  delay\_cause\_external\_cause  delay\_cause railinfrastracture  delaycause\_traffic\_management  delay\_cause\_rolling\_stock | Cluster Anlaysis (K-means) | By utilizing cluster analysis, we can group train stations based on their delay patterns and determine if specific train routes share similar delay characteristics. |
| 6.Can we predict the likelihood of train cancellations at the various stations based on historical data and external factors? | departure\_station, month, num\_of\_canceled\_trains, comment\_cancellations, delay\_cause\_external\_cause, delay\_cause\_rail\_infrastructure, delay\_cause\_traffic\_management, delay\_cause\_rolling\_stock, delay\_cause\_station\_management, delay\_cause\_travelers | Classification | Classification algorithms such as logistic regression, decision trees, random forests, or support vector machines can be used to predict the likelihood of delays for different train routes based on historical data and associated factors. |
| 7.Can we predict which train routes are likely to experience delays based on historical data? | departure\_station, arrival\_station, avg\_delay\_all\_departing, delay\_cause\_external\_cause, delay\_cause\_rail\_infrastructure, delay\_cause\_traffic\_management, delay\_cause\_rolling\_stock, delay\_cause\_station\_management, delay\_cause\_travelers | Classification | Classification algorithms such as logistic regression, decision trees, random forests, or support vector machines can be used to predict the likelihood of delays for different train routes based on historical data and associated factors. |
| 8.Can we forecast the average delay time at each station for the next month based on historical data? | departure\_station, month | Autoregressive Integrated Moving Average | Time series forecasting methods such as ARIMA (Autoregressive Integrated Moving Average), exponential smoothing, or state space models can be applied to forecast the average delay time at each station for the next month based on historical data. |

# Findings

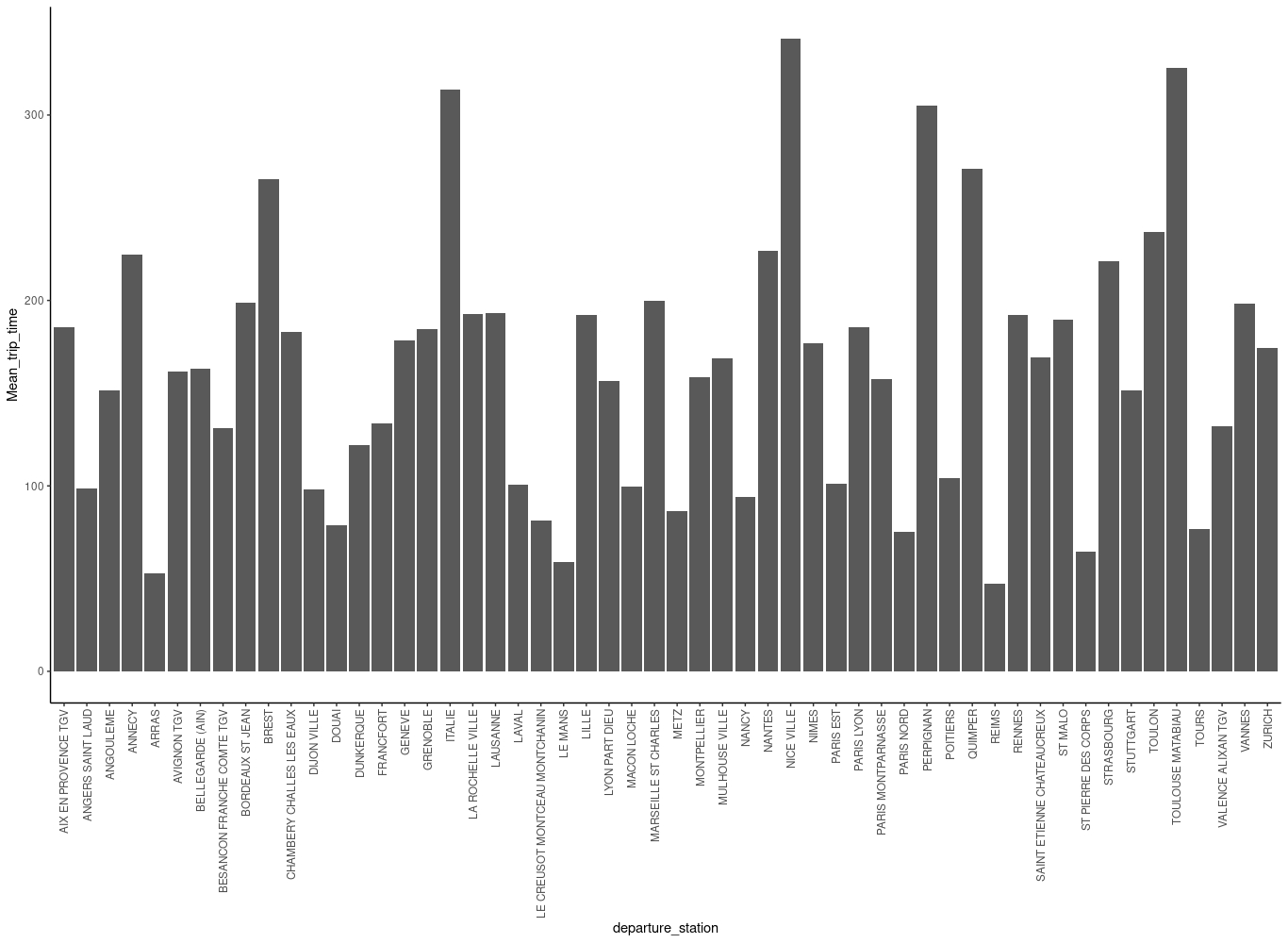
The distribution of canceled trains is majorly influenced by external factors and delays caused by infrastructure and delays by the travelers

Average trip times are majorly influenced by delays caused by management delays and rolling stock delays

variance in trips by month place PARIS LYON as the best performing station while ITALIE is the worst performing station.   
  
PARIS LYON and PARIS MONTPARNASSE have the highest number of trips and also have the highest number of canceled trips and delays.

# Plots

Distribution of trip times by stations

.

Distribution of cancelled trips by stations

